Norine: a powerful resource for novel nonribosomal peptide discovery

Maude Pupin, Qassim Esmaeel, Areski Flissi, Yoann Dufresne, Philippe Jacques, Valérie Leclère

Accepted Manuscript

Title: Norine : a powerful resource for novel nonribosomal peptide discovery

Author: Pupin M., Esmaeel Q., Flissi A., Dufresne Y., Jacques P., Leclère V.

PII: S2405-805X(15)30014-4
DOI: http://dx.doi.org/doi: 10.1016/j.synbio.2015.11.001
Reference: SYNBIO 1

To appear in: *Synthetic and Systems Biotechnology*

Received date: 15-10-2015
Revised date: 26-11-2015
Accepted date: 26-11-2015

Please cite this article as: Pupin M., Esmaeel Q., Flissi A., Dufresne Y., Jacques P., Leclère V., Norine : a powerful resource for novel nonribosomal peptide discovery, *Synthetic and Systems Biotechnology* (2015), http://dx.doi.org/doi: 10.1016/j.synbio.2015.11.001.

Norine : a powerful resource for novel nonribosomal peptide discovery

Pupin[1,2], M., Esmaeel[3], Q., Flissi[1,2], A., Dufresne[1,2], Y., Jacques[3,4], P. and Leclère[1,2,3,❖], V.

[1] : Univ Lille 1, CRIStAL, UMR CNRS 9189, F-59655 Villeneuve d'Ascq Cedex, France

[2] : Inria-Lille Nord Europe, F-59655 Villeneuve d'Ascq Cedex, France

[3] : Univ Lille 1, ICV Charles Viollette Institute EA 7394, ProBioGEM, Polytech'Lille, Avenue Langevin, F-59655 Villeneuve d'Ascq Cedex, France

[4] : Bioindustry Unit, Gembloux Agro-Bio Tech-University of Liege, Passage des Déportés, 2, 5030 Gembloux, Belgium

❖: corresponding author

Tel: +33320 43 46 68

Mail : valerie.leclere@univ-lille1.fr

Abstract

Since its first release in 2008, Norine remains the unique resource completely devoted to nonribosomal peptides (NRPs). They are very attractive microbial secondary metabolites, displaying a remarkable diversity of structure and functions. Norine (http://bioinfo.lifl.fr/NRP) includes a database now containing more than 1160 annotated peptides and user-friendly interfaces enabling the querying of the database, through the annotations or the structure of the peptides. Dedicated tools are associated for structural comparison of the compounds and prediction of their biological activities. In this paper, we start by describing the knowledgebase and the dedicated tools. We then present some user cases to show how useful Norine is for the discovery of novel nonribosomal peptides.

Keywords : Norine, nonribosomal peptides, secondary metabolites, database

Introduction

Nonribosomal peptides (NRPs) are attractive natural compounds because of their numerous biological activities potentially exploited by industries in diverse areas such as phytosanitary sector, cosmetics or health. They are produced by microorganisms (including bacteria and fungi) through specialized biosynthetic pathways. NRPs are biosynthesized by enzymatic modular complexes called NonRibosomal Peptide Synthetases (NRPSs) working as multidomain assembly lines[1]. The mode of synthesis leads to the production of compounds displaying a broad range of structures. Indeed, if some of them look like classical peptides because they are linear, most of them are more complex, including one or more cycles and branches. Moreover, those peptides are composed of monomers that are not limited to the 20 proteinogenic amino acids. Up to now, we have identified more than 530 building blocks

composing the different NRPs. The structural biodiversity is also due to the monomer modifications occurring during the synthesis made by the NRPSs themselves or performed post synthesis by accessory enzymes (also named tailoring or decorating enzymes). Famous examples for NRPs are the antibiotics penicillin[2], bacitracin and vancomycin[3], or the immunosuppressor cyclosporine[4]. In addition, some NRPs show antitumor activity such as Dactinomycin[5]. A current worrying public health issue is to find and develop new drugs to overcome multi-resistant pathogens. Therefore, it is important to develop bioinformatics tools for secondary metabolite discovery, such as antiSMASH[6] and tools especially dedicated to NRPSs and NRPs, such as Florine[7], NaPDos[8], and Norine[9,10]. The development of Norine was first motivated by the availability of computational tools allowing structure comparison of all NRPs[11], in spite of their complexity. For this purpose, we needed a database gathering all known NRPs, annotated according to their monomeric structure (i.e. monomer composition and 2D topology). Until now, the Norine team screened the literature to enter new peptides and annotated them manually. To get a more complete database, we have recently opened it to crowdsourcing through an easy-to-use web-based application[10]. Moreover, a semi-automatic process to extract data from external sources is currently under development.

The Norine database

Description and querying

Norine is a platform that includes the unique database dedicated to NRPs, associated with computational tools for their analysis. It has gained an international recognition thanks to high quality and manually curated annotations. Containing about 700 annotated NRPs for its first release in 2008, Norine database now contains more than 1160 NRPs that are clustered into 214 families, and composed of at least 530 distinct monomers. Among them, 73,5% have the "curated" status, which means that their nonribosomal origin is supported experimentally (for example due to identified NRPS) *vs* 26,5% are annotated with "putative" status due to

only presumed nonribosomal origin (often based on structural features). Two thirds of the peptides are cyclic or partially cyclic or contain at least one cycle. The sizes of the peptides range from 2 to 26 monomers, if polytheonamide is excluded, which was described as being the biggest NRP with 49 monomers for a long time but recently was identified to be a RiPP (Ribosomally synthesized and post-translationally modified peptide)[12]. Thus, in the near future, a third category will be created to tag all deprecated peptides when the hypothetical NRPS origin is finally excluded.

Each peptide page includes a comprehensive description of the peptide with the name, activities and structural atomic and monomeric details. The monomeric structure can be automatically obtained through the integrated smiles2monomers tool (s2m) when SMILES are available[13]. When identified, links to UniProt (for synthetases), PDB and PubChem (for structural data on the peptides) are provided. Moreover, a direct link to the NRPS gene clusters annotated in MIBiG[14] will be added soon.

Norine is queried from all over the world by biologists and biochemists to further analyze the nonribosomal peptides they study. For example, Desriac *et al.* [15] queried Norine to predict the antibacterial activity of a putative NRP produced by *Pseudoalteromonas*, while Bills *et al.*[16] used Norine to investigated the structural differences between bacterial and fungal NRPs. Indeed, for this purpose, the Norine platform provides visualization and editing applets for monomeric structure as well as tools to compare monomeric structures. Currently, Norine can be queried either by annotations (through "general search" tab) or by structural information (through "structure search" tab) of the peptides.

General search

Norine provides a basic interface that enables to query the database and search for peptides by combining multiple criteria, such as the name of the peptide, the Norine ID, the biological activities, the structure type, the producing organism, or the title or authors of references

associated to the NRP. The main advantage of this interface is that it allows users to extract data and get statisticss according to different criteria. For example, one can query for all siderophores produced by "any bacteria" (check "siderophore" in the "activity" field and enter "bacteria" in the "organism search" field), or all peptides with a linear structure, or simply search for all NRPs produced by the genus "*Pseudomonas"* (enter "pseudomonas" in the "organism search" field) (see the results in Fig. 1). The first output is a list of all the peptides corresponding to the criteria selected, classified by families. A click on a peptide name directs to the peptide page containing all details on the compound. Moreover, a click on the pie chart icon located above the list of results provides graphical output (Fig. 1). Pie charts and diagrams enable to filter the obtained results in order to refine them, by clicking on a slice, for example by structure type or monomers size.

Structure search

In addition to search through annotations, Norine proposes efficient structure search tools based on different algorithms: monomeric composition fingerprint (MCFP)[17], structure-based search for pattern comparison and similarity-based search[11,18]. These enable to find peptides containing a given list of monomers or a given 2D-pattern for structural comparisons. In Norine, a specific syntax is used (the NOR format) to describe the two-dimensions graph of a NRP, taking into account the topology of the molecule (linear, cyclic, branched, etc.). In the string representation used by the computational tools (i.e. "Val,Orn,Leu,D-Phe,Pro@1@0,2@1,3@2,4@3"), the monomers are listed, separated by commas; the @ character symbolizes the links between numbered monomers (explained in more details in the "structure search" part of the help tab). However, in most use-cases the string can be automatically generated with the graphical editor applet provided in the structure search form. Users only have to draw a peptide (or fragments of a peptide) by picking the monomers in the list proposed on the left side, and connecting the monomers, once they are placed on the

drawing area. Figure 2 illustrates an example of structure search results obtained for the linear peptide "Val_Orn_Leu_D-Phe_Pro" using the representation in NOR format generated by the graphical editor.

Advantages of structure-based search are manifold. First, using the graph representation of a NRP in Norine structure-based search tools enables to find similar NRPs that can be variants, or identical compounds from a structure point of view independent of their names / annotations. Second, we are convinced that the diversity of the biological activities of NRPs comes from their monomeric composition and the diversity of their structures[18]. That is the reason why the structure-based search tools can help to predict biological activities of an NRP. Indeed, similar NRPs probably share common properties such as their known activities. Finally, the structure comparison tools may be helpful to annotate NRPS coding genes/clusters within a microorganism genome sequence. Indeed, peptides with similar monomeric structures may be produced by NRPSs with close modular organization.

Examples presented below further illustrate the different use cases.

Norine database is now open to crowdsourcing

With the development of high throughput technologies dedicated to the screening of secondary metabolites, the number of published descriptions of new NRPs is increasing exponentially. Considering that researchers are the best experts to annotate the NRPs they are working on, we have decided to open Norine to crowdsourcing[10]. We have facilitated the process of entering NRPs into the database by developing an interface for peptide submission and modification by contributors/curators. In order to use this MyNorine interface, users firstly register by creating an account. Standardized forms are provided to submit new peptides or update records for existing peptide entries. The entered data are thereafter reviewed by validators of the Norine team. That process is crucial to ensure that peptides

stored in Norine are expert validated as this guarantees the quality of the data. The contributions will help to enrich the Norine database. The contributors will be mentioned as the authors of the entry.

Use cases

Identification of novel CLPs produced by Pseudomonas CMR12a

With the aim of discovering new cyclic lipopeptides (CLPs) with potential biocontrol activity, a combination of chemical structure analysis and *in silico* analysis of the genes encoding NRPSs was carried out on *Pseudomonas* CMR12a[19]. The strain was shown to produce two components originally named CLP1 and CLP2 with18 and 10 amino acid monomers within the peptide backbone, respectively. The structures of both compounds were elucidated and compared to the structure of all the peptides stored in Norine, using the "structure search" interface. A peptide named orfamide B was identified in the database that matched exactly to the peptide sequence of CLP2 (Fig. 3). CLP1 was identified as being a new member of the tolaasin group[20], displaying only one substitution on the monomer at position 6 (Fig. 3). Thus, the tolaasin group, which comprises at least 11 CLPs produced by different *Pseudomonas* strains (7 tolaasins, 2 corpeptins and 2 fuscopeptins), was extended with this new member, named sessilin according to its involvement in biofilm formation[19]. This example demonstrates the relevancy of structure search tool of the Norine resource to evaluate the novelty of peptides detected during a screening for active secondary metabolites.

Gratisin shares a pattern with the well-known gramicidin S

Gratisin is an undecapeptide with antibiotic activity produced by *Brevibacillus brevis* formerly named *Bacillus brevis* Y-33. Considering its primary cyclic structure[21], including a D-enantiomer of phenylalanin, and the presence of the nonproteogenic amino-acid ornithine,

it was assumed to be nonribosomally synthesized. An entry was created in the Norine database with a "putative" status because no NRPS associated with gratisin biosynthesis was known. The structure-based search returned 5 peptides sharing a pattern constituted of the pentapeptide motif "Val,Orn, Leu, D-Phe, Pro". All the 5 peptides display antibiotic activity and are produced by *Bacillus* strains: four belonging to tyrocidin family (tyrocidins A, B, C and D) and the fifth one is the decapeptide gramicidin S (Fig 2), in which the same pattern is repeated twice due to an iterative mode of biosynthesis[1,22] (Fig. 4). Even if the synthetase has not yet been identified in any sequenced *Bacillus* genome, we can guess that gratisin is nonribosomally synthesized with an NRPS also using an iterative mode of biosynthesis because it also contains a repeated motif. This example shows that the structure comparison of the final peptide may give insights into the biosynthetic pathway and thus contributes to the prediction of the modular organization of its producing NRPS. This can facilitate the identification of the genes or clusters directly involved in the production of such metabolites using genome-mining approaches.

Cepaciachelin : from putative to curated status

Twenty years ago, the structure of a siderophore produced by *Burkholderia ambifaria* strain PHP7 (LMG 11351) was elucidated[23]. It is a small compound composed of only 4 monomers: one lysine is bond to one putrescin and to two residues of di-hydoxy-benzoic acid (abbreviated Dhb or diOH-Bz). For siderophores a nonribosomal origin can not be systematically attributed because some of them, like anguibactin or enterobactin[24] are synthesized by NRPSs, whereas others, like desferrioxamin, are built up by other enzymes[25]. During a genome-mining analysis, we have identified a gene cluster within the genome of *B. ambifaria* AMMD that is responsible for the cepaciachelin production (personal

communication). Norine was directly queried with the NRP sequence predicted by antiSMASH, resulting in a hit against cepaciachelin. As there is a functional confirmation that cepaciachelin is indeed a NRP, the status Norine entry now could be updated from "putative" to "curated" (Norine ID = NOR01254). Cepaciachelin represents the fifth curated diOH-Bz containing peptide with a siderophore activity annotated in the Norine database.

Conclusion

Norine (http://bioinfo.lifl.fr/NRP) is a freely available and unique resource dedicated to nonribosomal peptides (NRPs)[10]. A user-friendly interface allows easy browsing, annotation, structure searching and downloading of the NRPs and their monomers. To discover new natural products, Norine may be the final step of a workflow, which is aimed at detecting the potential for new NRP biosynthesis from genomic data[7]. In the case where compounds are identified as being new NRPs or variants of an existing family, researchers can now submit them directly into Norine with the easy-to-use MyNorine interface. The scientific community will contribute to and benefit from the enriched resource, improving the screening for NRPs with biological or medical applications.

Acknowledgments

The authors are grateful to University of Lille 1, Inria-Lille Nord Europe, and bilille plateform for their financial support

References

1- Hur GH, Vickery CR, Burkart MD. Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. Nat Prod Rep 2012; 29: 1074-98.
2- Queener S. Molecular biology of penicillin and cephalosporin biosynthesis. Antimicrobiol agents chemother 1990; 34: 943-8.

3- Hubbard BK, Walsh CT. Vancomycin assembly: Nature's ways. Angew Chem Int Ed 2003; 42: 730-65.

4- Liu H, Wang Y, Li S. Advanced delivery of ciclosporin A : present state and perspective. Expert Opin Drug Deliv 2007; 44: 349-58.

5- Li JJ, Huang HH, Shen J, Jiang JY, Pan F, Yu ST, Kang Y, Zhao XX, Yan XC, Liang HJ. Multimodal therapy for adult Wilms'tumor : an experience from one centre. Clin Transl Oncol 2011; 13:672-6.

6- Weber T, Blin K, Duddela S, Krug D, Uk Kim H, Bruccoleri R, Yup Lee S, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E, Medema MH. antiSMASH 3.0 – a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucl Acids Res 2015; 43: W237-43.

7- Caradec T, Pupin M, Vanvlassenbroeck A, Devignes MD, Smaïl-Tabbone M, Jacques P, Leclère V. Prediction of monomer isomery in Florine: a workflow dedicated to nonribosomal peptide discovery. PLoS ONE 2014; 9: e85667.

8- Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR. The natural product domain seeker NaPDoS: a phylogeny based bioinformatics tool to classify secondary metabolite gene diversity. PloS ONE 2012; 7: e34064.

9- Caboche S, Pupin M, Leclère V, Fontaine A, Jacques P, Kucherov G. NORINE : a database of nonribosomal peptides. Nucl Acids Res 2008 ; D326-31.

10- Flissi A, Dufresne Y, Michalik J, Tonon L, Janot S, Noé L, Jacques P, Leclère V, Pupin M. Norine, the knowledgebase dedicated to nonribosomal peptides, is now open to crowdsourcing. Nucl Acids Res 2015; doi 10.1093lnarlgkv 1143.

11- Caboche S, Pupin M, Leclère V, Jacques P, Kucherov G. Structural pattern matching of nonribosomal peptides. BMC Structural Biology 2009; 9: 15.

12- Arnison PG *et al*. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommandations for a universal nomenclature. Nat Prod Rep 2013; 30:108-60.

13- Dufresne Y, Noé L, Leclère V, Pupin M. Smiles2Monomers : a link between chemical and biological structures for polymers. J Cheminform 2016 *in Press*.

14- Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, de Bruijn I, Chooi YH, Claesen J, Coates RC *et al*. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol 2015; 11: 625-31.

15- Desriac F, Jégou C, Balnois E, Brillet B, Le Chevalier P, Fleury Y. Antimicrobial peptides from marine proteobacteria. Mar Drugs 2013; 11 : 3632-60.

16- Bills G, Li Y, Chen L, Yue Q, Niu XM, An Z. New insights into the echinocandins and other fungal non-ribosomal peptides and peptaibiotics. Nat Prod Rep; 31: 1348-75.

17- Abdo A, Caboche S, Leclère V, Jacques P, Pupin M. A new fingerprint to predict nonribosomal peptides activity. J Comput Aided Des 2012; 26: 1187-94.

18- Caboche S, Leclère V, Pupin M, Kucherov G, Jacques P. Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. J Bacteriol 2010; 192: 5143-50.

19- D'aes J, Kieu NP, Leclère V, Tokarski C, Olorunleke FE, De Maeyer K, Jacques P, Höfte M. Ongena M. To settle or to move ? The interplay between two classes of cyclic lipopeptides in the biocontrol strain *Pseudomonas* CMR12a. Environm Microbiol 2014; 16:2282-300.

20- Roongsawang N, Washio K, Morikawa M. Diversity of nonribosomal peptide synthetases involved in the biosynthesis of lipopeptide biosurfactants. Int J Mol Sci 2011; 12:141-72.

21- Tamaki M, Takimoto M, Sofuku S,Muramatsu I. Synthetic studies on gratisin. J Antibiot 1983; 6:751-2.

22- Hoyer KM, Mahler C, Marahiel M. The iterative gramicidin S thioesterase catalyzes peptide ligation and cyclization. Chem Biol 2007; 14:13-22.

23- Barelmann I, Meyer JM, Taraz K, Budzikiewicz H. Cepaciachelin, a new catecholate siderophore from *Burkholderia* (*Pseudomonas*) *cepacia*. Z Naturforsch 1996; 51C: 627-30.

24- Crosa JH, Walsh CT. Genetics and assembly line enzymology of siderophores biosynthesis in bacteria. Microbiol Molec Biol Rev 2002; 66: 223-49.

25- Challis GL. A widely distributed bacteria pathway for siderophore biosynthesis independent of nonribosomal peptide synthetases. Chem Bio Chem 2005; 6: 601-11.

Figure legends

Fig. 1 : Graphical output provided with *Pseudomonas* query in "organism search" form

A) Pie charts representing the percentages of the nonribosomal peptides produced by *Pseudomonas*, according to their status, their class, their structure types and their activities.
B) Histogram representing size distribution of the peptides produced by *Pseudomonas*. For lipopeptides, the fatty acid is considered as one monomer.

Fig. 2 : Structure search results

Screenshots of the results obtained using the linear pentapeptide "Val_Orn_Leu_D-Phe_Pro" as a pattern (drawn as graph and transformed in NOR format by the graphical editor) for structural comparison.

Fig. 3 : Structural comparison of sessilin and CLP2 produced by *Pseudomonas* CMR12a with tolaasin and orfamide

The monomers have been aligned, the full lines represent the cyclization within the peptidic part. The variable amino acids in the peptide moiety are in bold and surrounded, small variability within the acyl moiety is highlighted by a dotted circle.

Fig. 4 : Structure comparison at the monomeric level

A) Gratisin, B) Gramicidin S. Top : monomeric representation as returned by the integrated smiles2monomers tool (s2m). Middle : monomers composing the peptides identified by s2m. Bottom : schematic representation highlighting the repetition of a common pentapeptide motif and the differences between both molecules.